

Sip or Savour?

Uncorking the Secrets of Wine Quality Prediction with Machine Learning

Dominik Zabinski 306068

Agenda

Sip or Savour?

Dominik Zabinski
306068

- ▶ Problem definition
- ▶ Data analysis
- ▶ Data preparation
- ▶ Methods
- ▶ Results
- ▶ Summary

Problem definition

- ▶ Which features are most important while predicting wine quality?
- ▶ What is the best model for this task?
- ▶ How good is the best model?

Data analysis

Sip or Savour?

Dominik Zabinski
306068

Distribution of wine quality

bins: observed, line: estimated

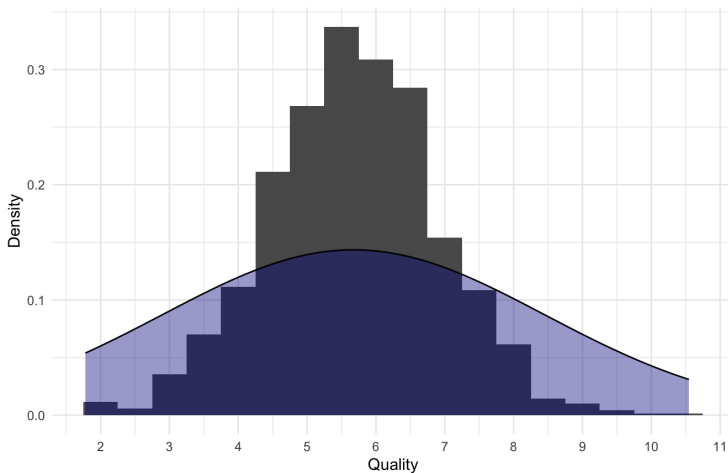


Figure 1: Wine quality distribution

Variable	Min	Avg	Sd	Max	# miss.	# dist.
alcohol	8.40	10.45	1.09	15.00	0	73
chlorides	0.00	0.09	0.05	0.61	0	184
citric.acid	0.00	0.27	0.19	1.00	0	75
density	0.98	1.00	0.01	1.01	0	676
feat01	0.00	0.51	0.29	1.00	0	1400
feat02	0.00	0.50	0.29	1.00	0	1400
feat03	0.00	0.50	0.29	1.00	0	1400
feat04	0.00	0.57	0.15	1.00	0	1400
feat05	0.00	0.50	0.29	1.00	0	1400
feat06	0.00	0.49	0.29	1.00	0	1400
feat07	0.00	0.53	0.16	1.00	0	1400
feat08	0.00	0.50	0.29	1.00	0	1400
feat09	0.00	0.48	0.28	1.00	0	1400
feat10	0.00	0.51	0.29	1.00	0	1400
fixed.acidity	4.80	8.47	1.70	16.10	0	92
free.sulfur.dioxide	3.00	17.68	10.18	74.00	0	56
pH	2.77	3.34	0.16	4.04	0	85
residual.sugar	0.80	2.48	1.50	15.40	0	84
sulphates	4.33	4.66	0.18	6.00	0	90
total.sulfur.dioxide	4.00	46.02	33.12	287.00	0	147
volatile.acidity	0.10	0.53	0.18	1.33	0	173

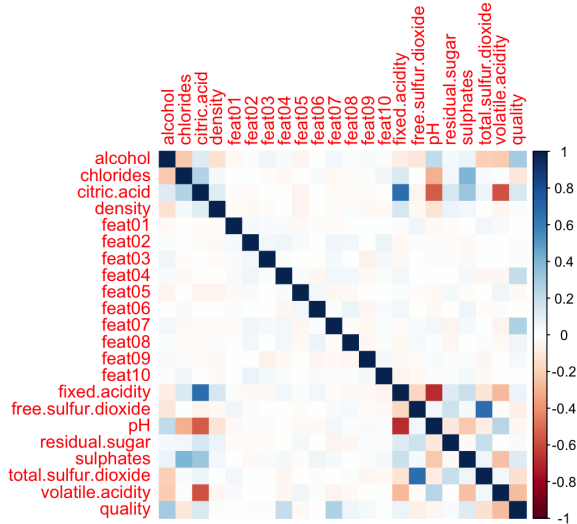


Figure 2: Correlation matrix between continuous variables

Data preparation

Sip or Savour?

Dominik Zabinski
306068

- ▶ split training/testing in 70/30 proportion
- ▶ normalize continuous variables
- ▶ 5-fold cross-validation as sampling method

Single models:

- ▶ Decision tree
- ▶ Random forest
- ▶ Generalized Boosted Regression Modeling (GBM)
- ▶ eXtreme Gradient Boosting (XGBoost)
- ▶ Linear regression (as benchmark)

Ensemble model: weighted combination of single models

Stacked models:

- ▶ Linear Regression as top layer model
- ▶ XGBoost as top layer model

Assessment metric: RMSE

Results

Sip or Savour?

Dominik Zabinski
306068

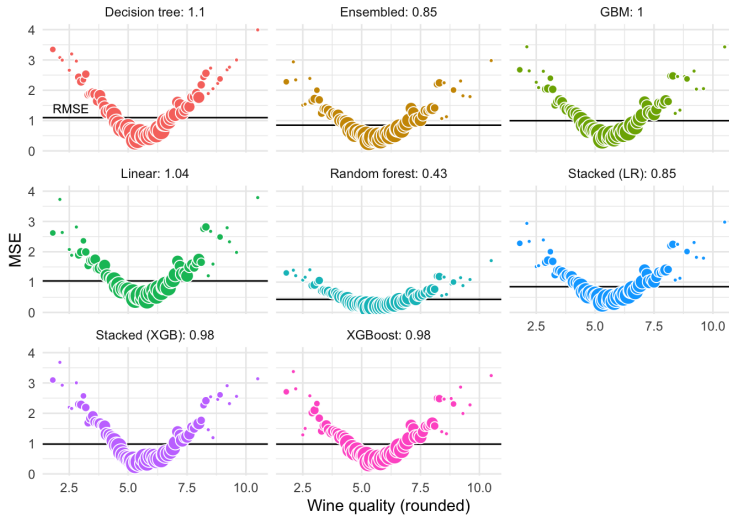


Figure 3: Model comparison - train data

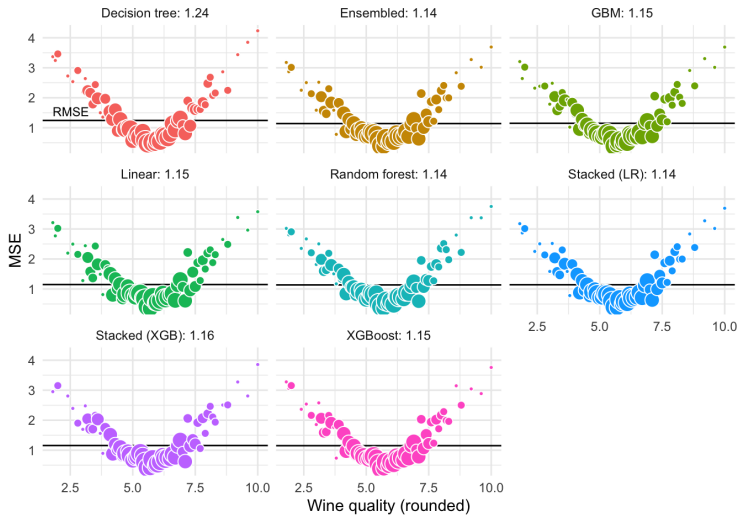


Figure 4: Model comparison - test data

Model	RMSE (Train)	RMSE (Test)
Linear	1.039	1.147
Decision tree	1.097	1.243
GBM	0.996	1.149
XGBoost	0.982	1.150
Stacked (XGB)	0.984	1.156
Ensembled	0.848	1.140
Stacked (LR)	0.848	1.140
Random forest	0.430	1.135

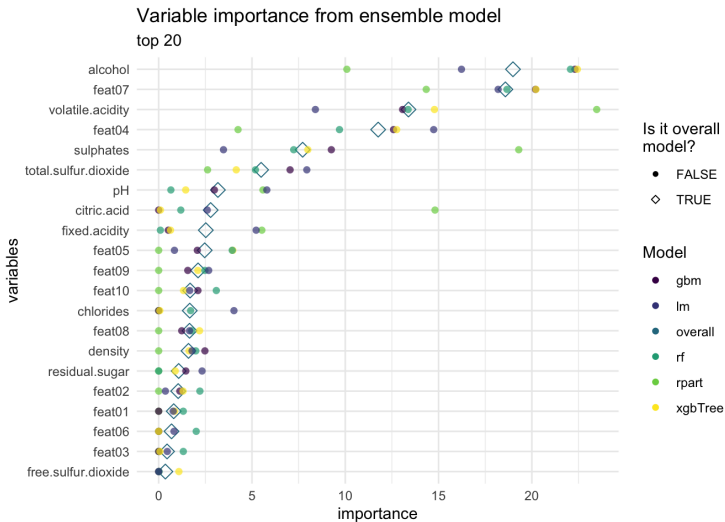


Figure 5: Variable importance - ensemble model

- ▶ Which features are most important while predicting wine quality?
 - ▶ alcohol, volatile acidity, sulphates
 - ▶ feat04 and feat 07
- ▶ What is the best model for this task?
 - ▶ Random Forest
- ▶ How good is the best model?
 - ▶ RMSE: 1.13
- ▶ What could be done differently?
 - ▶ hyperparameter tuning might improve the results
 - ▶ binning continuous variables